Data Pedigree in the Collaboratory for Multi-scale Chemical Science (CMCS)

Carmen Pancerella

carmen@ca.sandia.gov

Sandia National Laboratories

June 2002

Sandia National Laboratories

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Argonne National Laboratory

Los Alamos NATIONAL LABORATORY

NIST National Institute of Standards and Technology

MIT Massachusetts Institute of Technology

Berkeley University of California

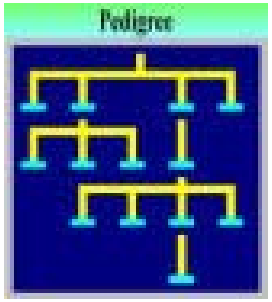
CMCS Pedigree

OBJECTIVE Data pedigree is a relationship which provides a "line of ancestors". This allows for the categorization and tracing of the scientific data, possibly across scales and for the identification of the data's ultimate origin. Pedigree data and metadata will be associated with CMCS resources, and will be browsable and searchable from the CMCS portal.

TASK This task involves both application scientists to define metadata and to determine what pedigree information is necessary and infrastructure team members to implement the solutions.

TEAM Carmen Pancerella (lead), Jim Myers, Brett Didier, Eric Stephan, John Hewson, Tom Allison, Branko Ruscic, David Leahy (honorary member), Michael Frenklach (honorary member)

Why is pedigree critical to CMCS?



- CMCS data will be identifiable and traceable.
- With a pedigree browser, a scientist can
  - trace scientific data to its ultimate origin, possibly across scales.
  - track data to program versions, and hence, to program bugs reported for that version.
  - retrieve data and get papers/reports about this data.
  - retrieve data and find references about its input data and parameters.
  - retrieve data at one scale and trace input data to techniques at a lower scale.

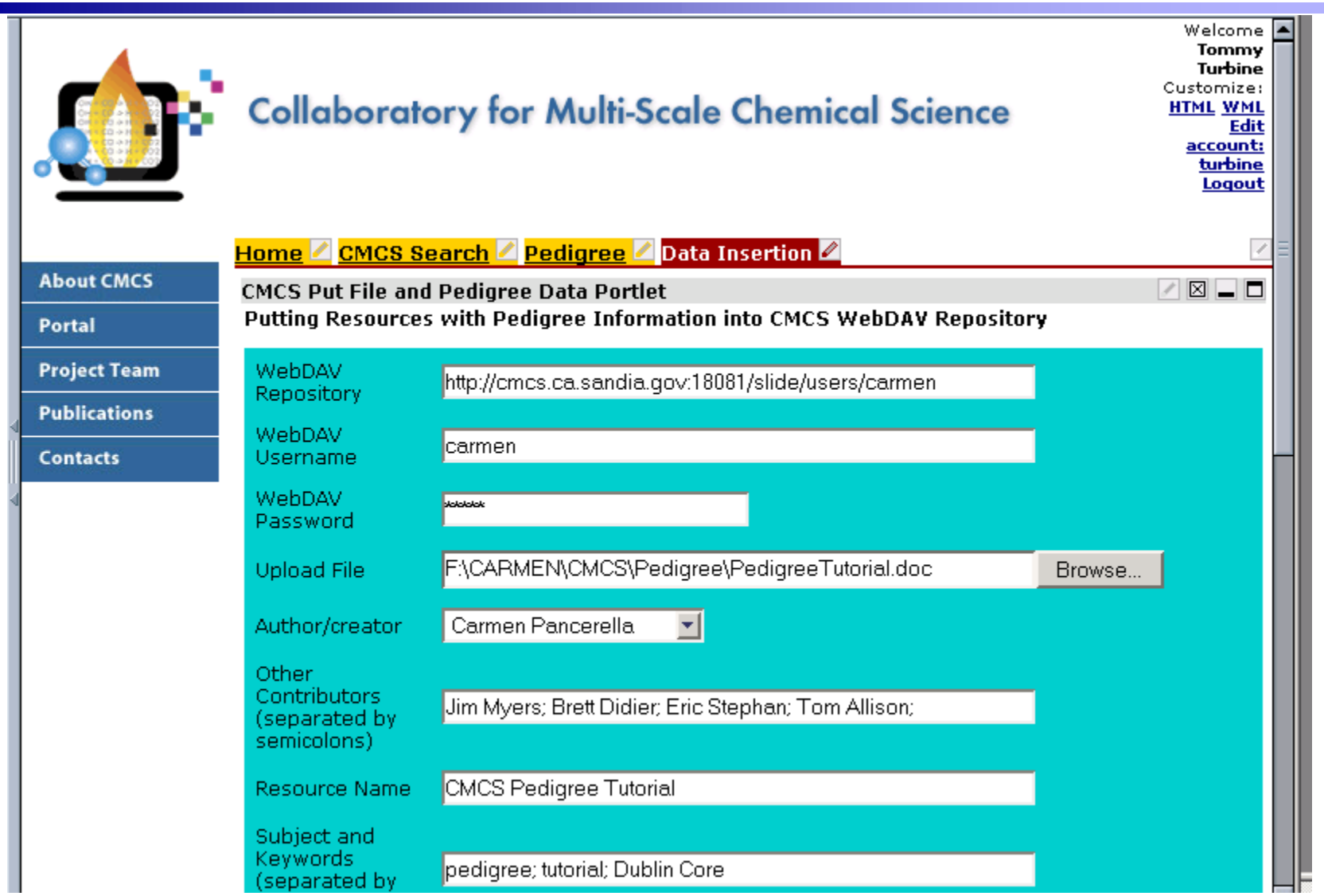
Pedigree Tasks

- [x] Document requirements for pedigree.
- [x] Create tutorial for CMCS team.
- [x] Survey metadata standards, in particular, Dubin Core. Make recommendations for use of Dublin Core in the CMCS.
- [ ] Define standard mechanism to represent subject-verb-object triples in DAV properties.
- [ ] Append data with pedigree information in WebDAV. At this time, GRIMech data (Michael Frenklach), HCT data (Bill Pitz), and EMSL Computation Results database (Theresa Windus) are on the Slide server and annotated with some pedigree metadata.

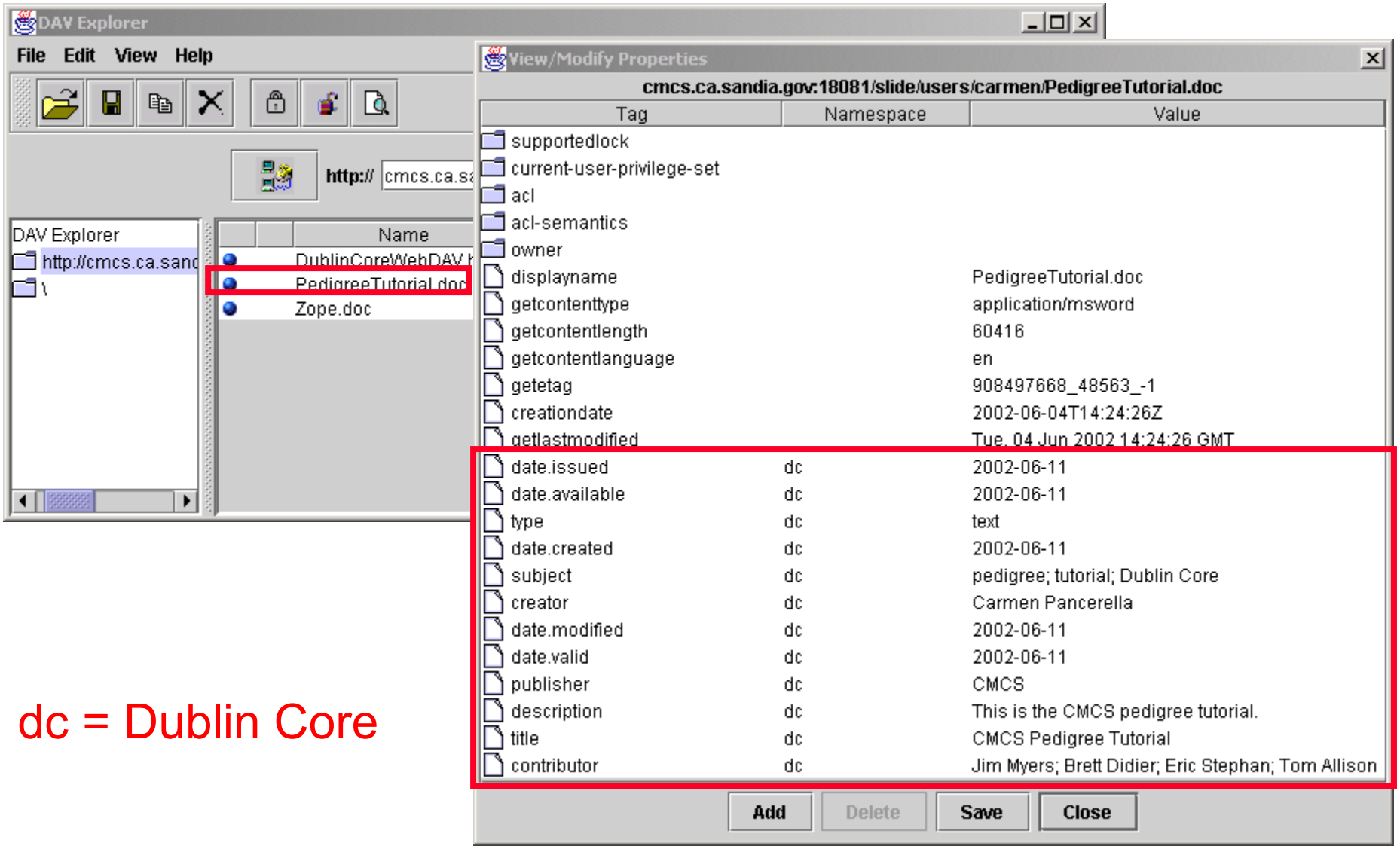
Pedigree Tasks

- Develop portlets to facilitate insertion and modification of pedigree metadata.
- Develop pedigree browser portlet.
- To Do List Develop pedigree API to facilitate insertion and modification of pedigree metadata.
- To Do List Develop pedigree API for pedigree browsing.

Portlet for Adding Resources and Pedigree Information Into CMCS

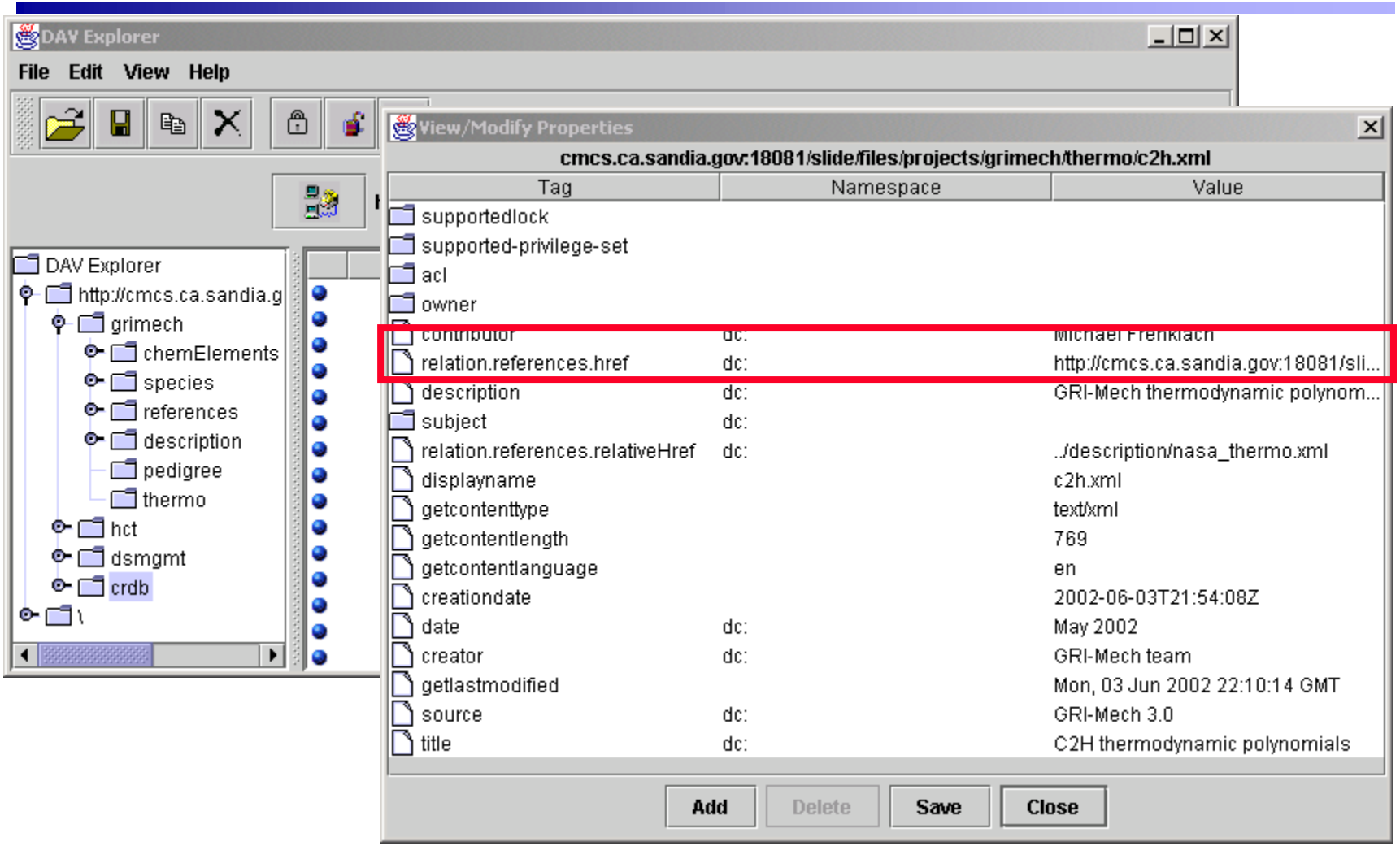


Pedigree Metadata Stored as WebDAV Properties



dc = Dublin Core

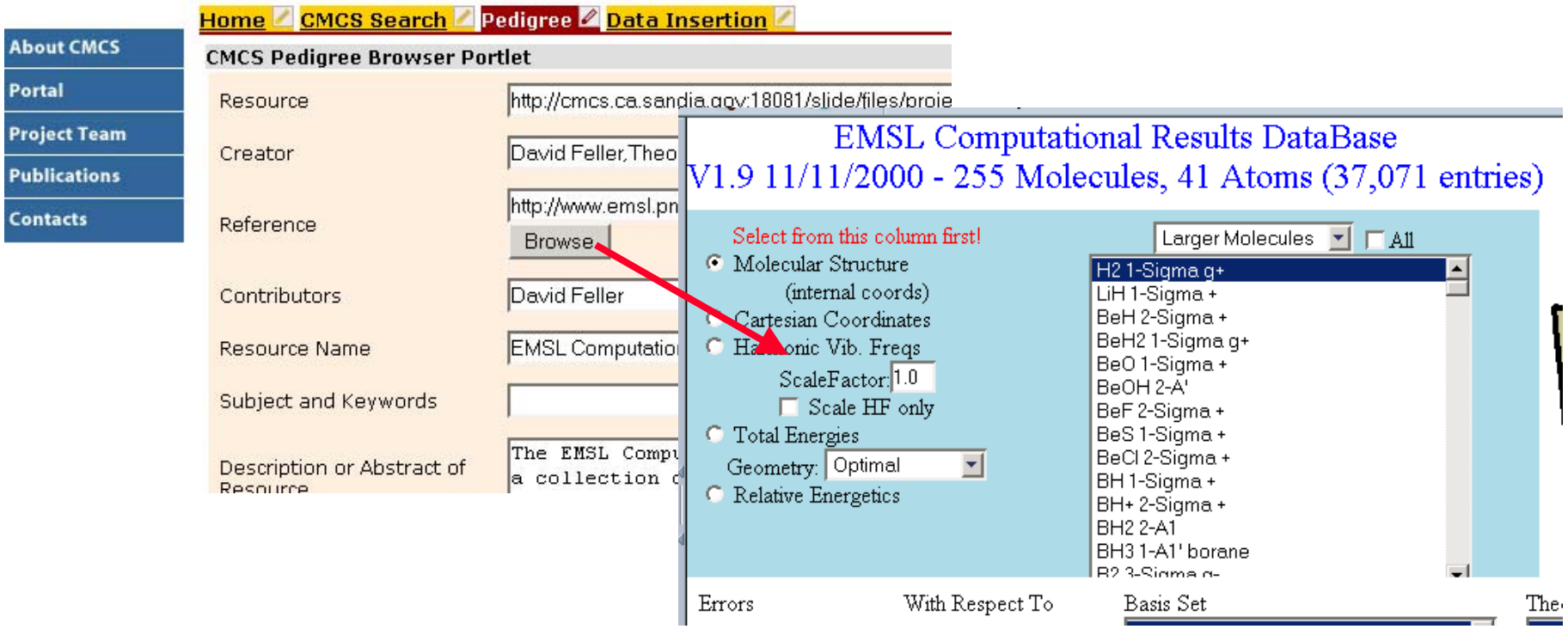
GRIMech Data on CMCS Slide Server Has Pedigree Metadata



GRIMech thermo data is annotated with pedigree information, including bibliographical references.

Pedigree Browser Portlet

Collaboratory for Multi-Scale Chemical Science



A reference URL is stored as pedigree metadata for the EMSL Computational Results Data on the CMCS Slide server. Pedigree browser allows this URL to be retrieved with a click.